KM.RU на РОМИП-2009. Получение стабильных результатов на разных коллекциях

© Сергей Татевосян, Наталья Брызгалова

«КМ онлайн»
{tatevosyan, bryzgalova}@post.km.ru

Аннотация

Работа посвящена участию проекта Поиск KM.RU в семинаре РОМИП-2009. В статье дается описание модификаций поискового алгоритма, рассматривается новый подход к оптимизации коэффициентов, представлены результаты участия KM.RU в дорожках поиска РОМИП-2009 и их анализ.

I. Введение

В 2009 году мы по-прежнему являемся участником поисковых дорожек на семинаре РОМИП, и главная цель нашего участия - проверка стабильности результатов на разных коллекциях документов. Мы исследовали эффективность модифицированного поискового алгоритма, в котором появились новые факторы и был улучшен способ учета старых. В то же время изменился принцип оптимизации параметров. Оптимизация проходила на базе оценок экспертов РОМИП, поэтому для нас участие в семинаре – это, помимо проверки работоспособности алгоритма, прекрасная возможность пополнить базу оцененных документов новыми запросами.

II. Особенности поискового алгоритма. Оптимизация коэффициентов

В этом году мы реализовали возможность поиска по кворуму. Применение кворума позволяет достигать бóльших значений полноты, однако при этом возникает риск снижения показателей точности для поисковой выдачи. В экспериментах РОМИП мы рассчитывали на увеличение значений показателя Recall. Нам было интересно, насколько хорошо алгоритм справится с задачей выбора лучших документов при заметном увеличении числа найденных.

Вторым принципиальным отличием алгоритма этого года стал новый метод оптимизации параметров. Тема оптимизации становится популярной, – к примеру, в этом году конкурс Яндекса «Интернет-математика» был целиком посвящен оптимизации - поскольку при постоянно растущем количестве параметров существует необходимость грамотного их использования.

Помимо указанных нововведений, мы добавили в основную формулу дополнительные параметры, часть из которых представляет собой модифицированный вариант «длинных» пассажей, предложенных в [3].

1. Поиск и ранжирование документов

1.1 Поисковая формула

Для определения степени соответствия документа запросу мы использовали следующую зависимость:

W = W док + W с + W пас + Wк + Wдоп (1),

где W – итоговое значение релевантности документа,

W док – вес документа, рассчитанный по TF*IDF модификации BM25 с учетом PageRank,

W с – вес ссылок на документ,

W пас – вес пассажей в документе,

W к – добавка за число слов из запроса в документе (за прохождение кворума),

W доп – дополнительные параметры: например, близость слов из запроса к началу предложения.

1.2 Новые параметры

В этом году главным нововведением для нас стало использование кворума при обработке запроса. Во время рабочих тестирований алгоритм с кворумом демонстрировал значительный прирост качества, поэтому нам важно было проверить работу кворума в рамках экспериментов для РОМИП. Кроме того, мы опробовали некоторые другие поисковые параметры.

а) кворум;

В новой поисковой платформе мы используем «мягкий» кворум, схожий с описанным в [4].

Первоначально принцип работы кворума был связан с данными о весах слов в запросе. Мы исходили из предположения: чем больше IDF, тем лучше слово. Например, при отработке кворума на запросе из 5 слов, когда достаточно взять всего три слова, мы брали те 3 слова, которые весят больше всего. Эксперимент по применению кворума с такими условиями дал ухудшение качество ранжирования.

В дальнейшем мы выяснили, почему это происходит, и наметили пути к повышению качества. Фактически, отрицательный результат дал толчок к новому подходу поиска релевантных документов.

Второй подход к вычислению кворума основывался на предположении, что веса всех слов в запросе следует считать одинаковыми. Такой подход дал положительные результаты: качество ранжирования при использовании кворума повысилось.

Проведя следующий цикл экспериментов, мы пришли к выводу, что смысл запроса может заключаться в словах как с самым большим, так и с самым маленьким весом. Например, в запросе «фильм район № 9» значимыми словами являются «район 9», а слово «фильм» можно опустить. При этом IDF слова *9* ниже, чем IDF слова *фильм*.

Особенности. Анализ запроса

Насколько мы заметили по результатам рабочих тестирований, одна из проблем в использовании кворума связана с запросами, содержащими «слова-указатели», которые совершенно необязательно буду присутствовать в релевантном документе. Это

запросы типа «фильм район № 9», «произведение Набокова Дар», «классификация породы кошек»: здесь слова *фильм*, *произведение*, *классификация* не несут большой смысловой нагрузки и могут даже не встретиться в релевантном документе, однако слова, которые как раз таки важны: *9*, *дар*, *кошек* – не являются такими уж редкими и весомыми, поэтому при отработке кворума возникает опасность, что кворум отсечет именно их (*9*, *дар*, *кошки*), вместо того чтобы отсечь *фильм*, *произведение*, *классификация*, и мы найдем документы со словами «фильм район», «произведение Набокова», «классификация породы».

Однако на данную проблему можно посмотреть и с другой стороны. Указанные выше слова, не несущие большой смысловой нагрузки для данного запроса, можно рассматривать как указатели на ветвь каталога или область знаний, которой, скорее всего, будет принадлежать релевантный документ. Таким образом, получив запрос «произведение Набокова Дар», мы можем сразу сосредоточиться на поиске по книжным Интернет-магазинам или Интернет-библиотекам, а получив запрос «сантехника Москва», видим географический указатель, который предполагает поиск по сайтам магазинов, продающих сантехнику в Москве. Такие «служебные» слова в запросе могут указывать на сам предмет поиска («текст романа Война и мир»), на действие, которое хочет совершить пользователь («Терминатор 4 скачать»), на географическую привязку («сантехника Москва»). По сути, слова-указатели – это служебная информация, которая совсем необязательно должна присутствовать в тексте документа, но может встречаться в ссылках на документ, в заголовке или в подразделах меню на странице, которые часто рассматриваются как не несущее смысл оформление. Современные поисковые системы в Интернете частично используют подобную информацию в запросе, показывая на запрос «карта N» карту города *N* над поисковой выдачей. Полагаем, что выделение и использование таких слов-указателей – возможный путь к структуризации коллекции документов и сужению области поиска релевантных страниц.

б) исправление опечаток;

В этом году в рамках РОМИП мы проверяли работу модуля исправления опечаток. Для этого запросы прогонялись через программу-корректор. На выходе получился список запросов с исправленными опечатками, и прогоны для РОМИП делались уже по новому списку.

Особенности

Автоматическое исправление опечаток – неоднозначный процесс, т.к. в каждом четвертом-пятом случае машинному алгоритму сложно выбрать единственный вариант исправления, поэтому в промышленных поисковых системах исправление опечаток начинают вводить, сначала предлагая пользователю самому выбрать правильный вариант в спорном случае. Это позволяет собрать дополнительные статистические данные для лучшей отработки алгоритма в будущем.

Однако в рамках экспериментов РОМИП нам нужно было в каждом случае выбрать только один верный вариант.

в) длинные пассажи

Идея «длинных» пассажей была предложена одним из участников РОМИП-2008 [3]. Смысл «длинных» пассажей заключается в том, что слова, недостающие для целого пассажа в теле документа, можно взять из заголовка или начала документа. Мы дополнили алгоритм возможностью использования слов из ссылок на документ. В рабочих экспериментах этот параметр не продемонстрировал прироста качества. Возможно, это объясняется отсутствием запросов, для которых релевантные документы имели бы соответствующие закономерности.

Помимо заголовка и ссылок, важные слова для пассажей на странице могут содержаться в тексте оформления, если оформление является вложенным меню на сайте. Например, на сайте Интернет-магазина, на странице со стиральными машинами может находиться цепочка вложенного меню Бытовая техника -> Стиральные машины -> Indesit. И для запроса «стиральная машина Indesit» страница с пассажем, составленным из текста оформления, может быть более подходящей, чем страница с полным пассажем, расположенным только в теле документа.

Особенности. Понятие «составных» запросов

В процессе рабочих исследований мы определили для себя понятие «составных» запросов.

Под «составным» запросом мы понимаем такой, в котором, помимо основного смысла, существует дополнительная служебная

информация. Например, запрос «поэма А.С. Пушкина Евгений Онегин»: для получения релевантных документов в выдаче будет достаточно словосочетания «Евгений Онегин», более того, остальные слова в запросе могут оказаться не только ненужными, но и ухудшить выдачу.

Современные методы поиска, говоря о пассаже в тексте документа, фактически ищут цитату. Но для данного запроса необходимым полным пассажем в тексте будет «Евгений Онегин», остальные слова из запроса могут по отдельности встретиться в заголовке документа, ссылках на него или не встретиться совсем. Таким образом, наилучший пассаж должен определяться не по IDF слов, входящих в него, а по степени смысловой значимости слов из пассажа для данного запроса. Уравновесить подобные противоречия частично помогают «длинные» пассажи: неполный пассаж в теле документа можно дополнить словами из заголовка документа и из ссылок на документ.

Стоит отметить, что «составные» запросы могут содержать не только служебную, но и просто избыточную, ненужную информацию. Поэтому наша задача – уметь выделять эти запросы среди других и вычленять из них различные виды информации.

При составлении запроса пользователи обычно употребляют:

1. Слова, несущие основную информацию, которые зачастую встречаются в документе в виде цитаты: например, словосочетание «Евгений Онегин» из запроса «поэма А.С. Пушкина Евгений Онегин»;
2. Слова со служебной информацией. Это могут быть указатели на область поиска основной информации: например, слово «скачать» в запросе «фильм терминатор 4 скачать» (т.е. релевантным будет документ не просто о фильме, а страница, где можно скачать искомый фильм), слово «Москва» в запросе «теннисные корты Москва»;
3. Слова с избыточной информацией, которая оказывается ненужной, а иногда может и мешать поиску: слово «поэма» из запроса «поэма Евгений Онегин», слово «почитать» из запроса «текст Библии почитать»;
4. Слова-синонимы или слова, по смыслу связанные с ключевыми. Например, вместо запроса «мебель для кухни» пользователь может сформулировать запрос, как «гарнитур для кухни», тогда как документов со словосочетанием «гарнитур для кухни» меньше, чем документов со словосочетанием «мебель для кухни». Для обработки

подобных запросов можно использовать тезаурус для выделенной тематической области или «базу связанных понятий» [5].

2. Оптимизация коэффициентов

Мы модифицировали способ оптимизации коэффициентов для слагаемых, входящих в формулу релевантности. Использование нового способа показало прирост качества в наших экспериментах. Способ оптимизации взят из [2].

Порядок оптимизации:

а) Выделяются группы взаимовлияющих друг на друга параметров. Мы это делали вручную, исходя из общих соображений. Оптимальным будет найти автоматический способ выделения групп таких параметров.

б) Сначала оптимизируется группа наиболее значимых параметров, потом менее и т.д.

Важно найти хорошее начальное приближение, так как при плохом функция сваливается в локальный минимум, из которого может не выбраться.

III. Участие в семинаре. Результаты

1. Эксперименты для РОМИП

1.1 Дорожки

В этом мы году приняли участие в дорожке поиска по веб-коллекции и в дорожке поиска по нормативно-правовой коллекции.

Как и в прошлом году, веб-коллекция состояла из коллекции BY – это выборка из страниц с домена .by, содержащихся в индексе поисковой системы Яндекс в 2007 году, и из коллекции KM, которая является копией мультипортала km.ru по состоянию на 2007 год. Участниками семинара было принято решение не вносить изменения в состав веб-коллекции, поскольку это позволит собрать больше материала из оцененных экспертами документов.

Нормативно-правовая коллекция состояла из документов Законодательства Российской Федерации, Москвы и Санкт-Петербурга.

1.2 Прогоны

Для дорожки поиска по нормативно-правовой коллекции мы сделали один прогон со своим базовым алгоритмом с использованием кворума при обработке запроса.

Для дорожки поиска по коллекции BY мы сделали прогон с нашим базовым алгоритмом с использованием кворума и с включенным механизмом исправления опечаток, коэффициенты были получены после оптимизации новым методом (см. Раздел II.2).

Для дорожки поиска по коллекции КМ мы сделали прогон со своим базовым алгоритмом с использованием кворума, новой оптимизацией и включенным механизмом исправления опечаток.

2. Результаты прогонов. Анализ результатов

Проанализировав полученные результаты, мы предположили, что для улучшения результатов по коллекции КМ можно смягчить условия кворума, а для коллекции BY – изменить параметры учета начала предложения.

Это дало прирост качества. Результаты с указанными модификациями (*KM-after*), полученные вне прогонов РОМИП, отображены на рисунках вместе с официальными графиками и показателями.

Коллекция BY

По коллекции BY мы получили результаты, представленные ниже (оценки OR и AND).

На рисунках можно выделить верхнюю группу графиков, которые демонстрируют схожий характер. У прогонов, которым соответствуют данные графики, другие оценочные показатели имеют близкие значения.

И при сильных, и при слабых требованиях к релевантности графики участников имеют одинаковый характер.

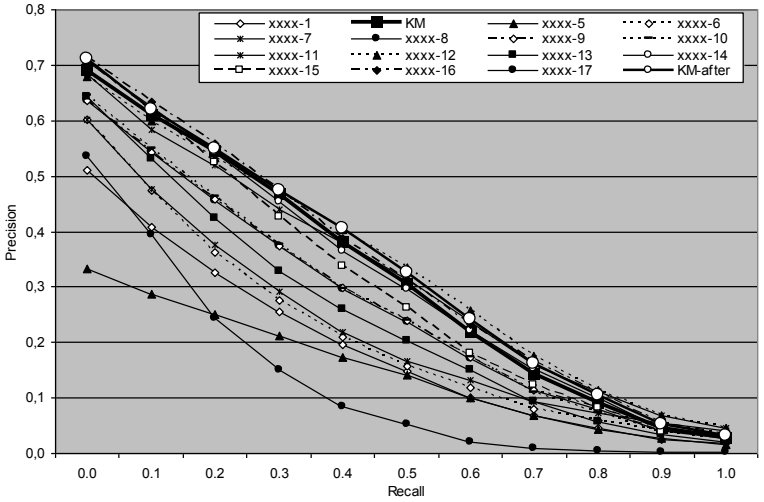

Рисунок 1. График TREC для коллекции BY.WEB, оценка OR

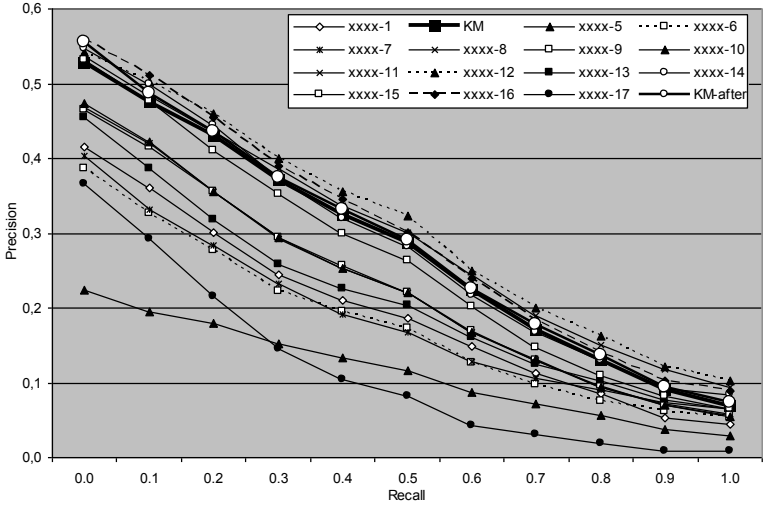

Рисунок 2. График TREC для коллекции BY.WEB, оценка AND

Ниже приведена сравнительная таблица для всех участников по основным показателям, оценка OR:

№	Prec(5)	Prec(10)	Bpref-10	Bpref	Recall
xxxx-1	0,36	0,33	0,22	0,20	0,28
KM	***0,52***	***0,48***	***0,38***	***0,34***	***0,54***
xxxx-5	0,25	0,22	0,18	0,16	0,25
xxxx-6	0,40	0,36	0,28	0,24	0,39
xxxx-7	0,40	0,37	0,28	0,24	0,40
xxxx-8	0,44	0,41	0,33	0,29	0,47
xxxx-9	0,44	0,41	0,33	0,29	0,47
xxxx-10	0,45	0,42	0,33	0,29	0,47
xxxx-11	0,51	0,46	0,37	0,33	0,51
xxxx-12	0,51	0,47	0,39	0,35	0,55
xxxx-13	0,45	0,40	0,30	0,27	0,43
xxxx-14	0,53	0,49	0,38	0,34	0,52
xxxx-15	0,52	0,47	0,36	0,32	0,50
xxxx-16	0,53	0,49	0,39	0,35	0,54
xxxx-17	0,35	0,32	0,16	0,15	0,20

В этом году основным нашим нововведением было использование кворума, поэтому мы рассчитывали, в частности, на улучшение показателя Recall по сравнению с прошлым годом. Наши расчеты в отношении этого параметра подтвердились.

Значение Recall, а также значения метрик Precision(5), Precision(10), Bpref-10, Bpref у нас на хорошем уровне и отличаются от показателей лидера всего на 0,01.

Коллекция KM

Графики для коллекции KM (оценки OR и AND) приведены ниже на Рисунке 3 и Рисунке 4.

При сравнении графиков TREC, построенных для оценки OR и AND, мы видим, что характеры графиков меняются в зависимости от оценок. Для графиков, построенных по результатам оценки OR, характерен одинаковый излом, все кривые идут практически параллельно друг другу. На рисунке с графиками, построенными по результатам оценки AND, картина совсем другая.

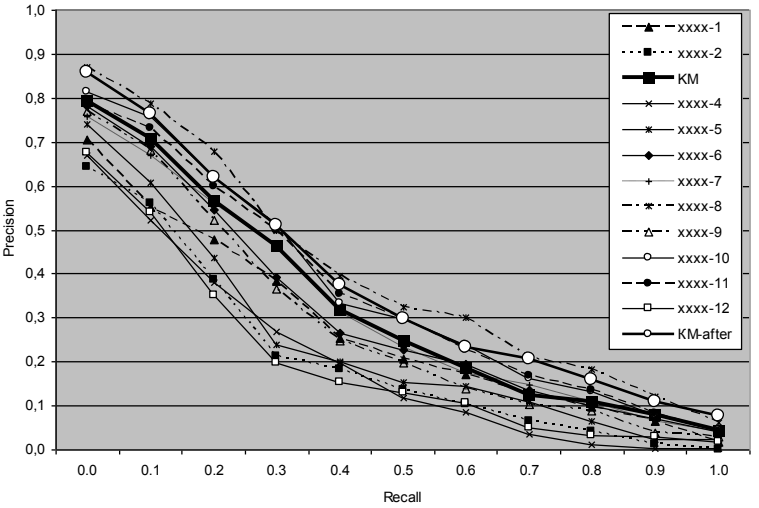

Рисунок 4. График TREC для коллекции KM.RU, оценка OR

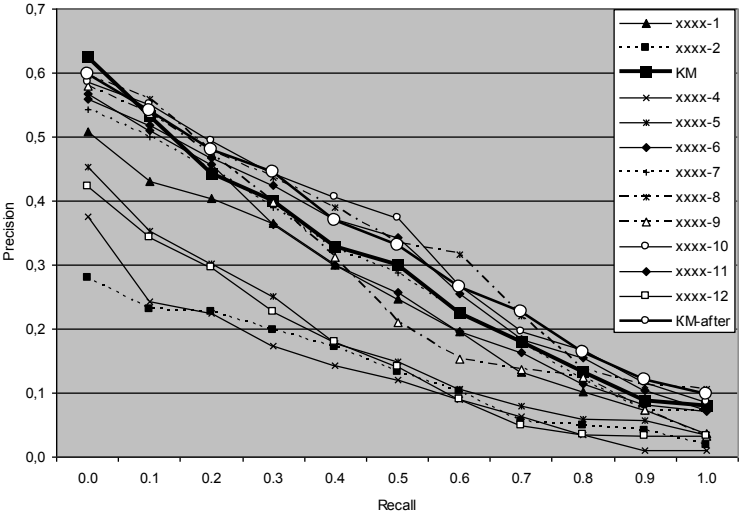

Рисунок 3. График TREC для коллекции KM.RU, оценка AND

Предсказуемо, что более строгий кворум дает лучшее начало графика и больший Prec(5) – график KM(AND), тогда как смягчение кворума существенно увеличивает Recall и поднимает график в

целом – KM-after(AND). Цель: при хорошей высоте графика - максимально высокая первая точка.

В следующей Таблице приведены результаты участников по метрикам Precision(5), Precision(10), Precision, Recall, оценка AND.

№	**Prec(5)**	**Prec(10)**	**Precision**	**Recall**
xxxx-1	0,34	0,30	0,17	0,43
xxxx-2	0,16	0,17	0,09	0,29
KM	***0,41***	***0,38***	***0,18***	***0,64***
xxxx-4	0,19	0,18	0,11	0,35
xxxx-5	0,24	0,25	0,12	0,37
xxxx-6	0,35	0,32	0,18	0,65
xxxx-7	0,33	0,33	0,18	0,63
xxxx-8	0,40	0,38	0,18	0,63
xxxx-9	0,40	0,39	0,16	0,49
xxxx-10	0,39	0,38	0,20	0,64
xxxx-11	0,38	0,39	0,20	0,64
xxxx-12	0,36	0,29	0,08	0,27
KM-after	*0,40*	*0,41*	*0,19*	-

Ниже - результаты участников по основным показателям, оценка OR:

№	**Prec(5)**	**Prec(10)**	**Precision**	**Recall**
xxxx-1	0,58	0,54	0,35	0,35
xxxx-2	0,48	0,49	0,26	0,30
KM	***0,59***	***0,59***	***0,37***	***0,47***
xxxx-4	0,49	0,48	0,28	0,33
xxxx-5	0,55	0,53	0,28	0,34
xxxx-6	0,63	0,57	0,33	0,46
xxxx-7	0,59	0,57	0,34	0,46
xxxx-8	0,74	0,68	0,39	0,54
xxxx-9	0,64	0,60	0,31	0,39
xxxx-10	0,67	0,59	0,38	0,49
xxxx-11	0,64	0,58	0,38	0,49
xxxx-12	0,57	0,54	0,22	0,25
KM-after	*0,67*	*0,63*	*0,37*	-

В оценках для коллекции KM нас также интересовал показатель Recall, который должен был подняться на хороший уровень за счет введения кворума. Наши ожидания оправдались.

Хотим отметить, что показатели для оценки AND оказались лучше, чем для OR: у нас лучшая среди участников оценка Prec(5), на хорошем уровне Prec(10), Precision, Recall. Мы связываем хорошие оценки AND с использованием ссылочного ранжирования, когда документы с высокой степенью доверия (имеющие высокий PageRank) оказываются выше остальных.

Коллекция Legal

В полученных результатах оценки дорожки поиска по Legal каждый документ оценен экспертом только один раз - поэтому результаты оценки не зависят от слабых или сильных требований к релевантности.

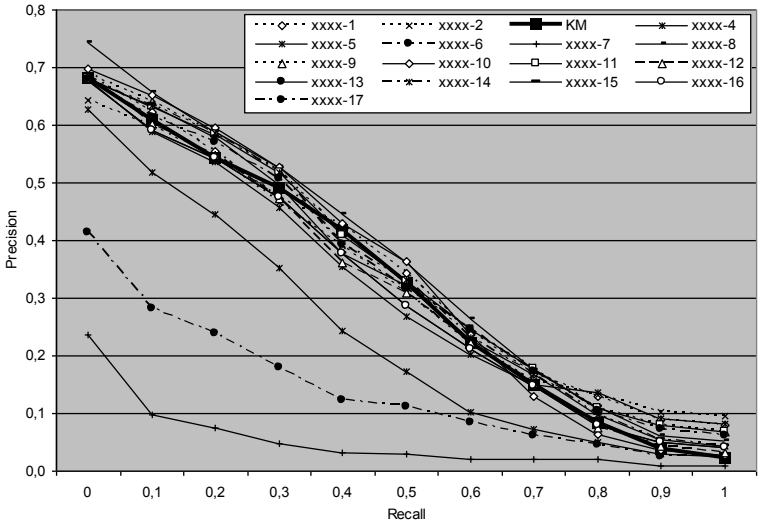

Рисунок 5. График TREC для коллекции Legal

В Таблице ниже приводим результаты участников по оценкам Precision(5), Precision(10), Precision, Recall:

№	Prec(5)	Prec(10)	Precision	Recall
xxxx-1	0,52	0,48	0,31	0,53
xxxx-2	0,51	0,47	0,27	0,50
КМ	***0,56***	***0,50***	***0,27***	***0,52***
xxxx-4	0,46	0,44	0,21	0,42
xxxx-5	0,43	0,41	0,21	0,47
xxxx-6	0,24	0,22	0,11	0,24
xxxx-7	0,09	0,08	0,05	0,14
xxxx-8	0,58	0,51	0,27	0,58
xxxx-9	0,57	0,51	0,24	0,52
xxxx-10	0,53	0,49	0,26	0,51
xxxx-11	0,55	0,51	0,24	0,51
xxxx-12	0,55	0,49	0,23	0,50
xxxx-13	0,54	0,50	0,23	0,51
xxxx-14	0,55	0,50	0,24	0,52
xxxx-15	0,55	0,50	0,24	0,51
xxxx-16	0,54	0,50	0,23	0,51
xxxx-17	0,52	0,50	0,24	0,52
xxxx-18	0,45	0,43	0,21	0,49

У нас хорошие результаты по всем метрикам. Значения Precision(5), Precision (10), Precision на уровне лидера.

IV. Потенциальные возможности совершенствования поискового алгоритма. Лингвистика, дискурсивный анализ

Все основные параметры нашего поискового алгоритма можно назвать статистическими, однако очевидно, что для совершенствования поискового механизма необходимо обращаться не только к математическим данным.

С одной стороны, значительно улучшить качество можно, только сосредоточившись на конкретной, а не абстрактной задаче информационного поиска. В семинаре РОМИП принимают участие проекты, перед которыми стоят разные задачи в реальной рабочей жизни, поэтому участие в поисковых дорожках РОМИП не всегда позволяет развернуться в полную силу. Так, одни системы, допустим, умеют производить сложную синтаксическую оценку запроса и предложений в документе, что может обеспечить им

лучшие результаты в экспериментах РОМИП. Такое не может себе позволить поисковая система в Интернете, для которой плюс одна секунда на ответ – это потеря части пользователей. Однако и наоборот, в поиске по веб-коллекции «научные» поисковые проекты могут показать себя не настолько блестяще, как в поиске по конкретной тематической области, для которой у них создан тезаурус, семантическая сеть или особый словарь замены тематических понятий.

С другой стороны, можно пробовать улучшить качество результатов поиска, привлекая для этого знания, которые дает языковой анализ материала. Безусловно, применение лингвистических методов в информационном поиске значительно «утяжеляет» алгоритм, и, если результаты станут лучше на 20-30 процентов, скорость обработки запроса может увеличиться на столько же или вообще в несколько раз. Однако поисковые системы в Интернете не должны скидывать языковой анализ со счетов. На самом деле, очень многие статистические параметры имеют под собой лингвистическую основу.

К примеру, учет близости слов из запроса к началу предложения обусловлен лингвистически: в лингвистике есть понятие темы и ремы предложения, где тема – это старая информация, которая уже известна, а рема – новая информация, которая только вводится в повествование, и в предложениях в русском языке они наиболее часто следуют в порядке «рема – тема». Таким образом, отдавая предпочтение словам в начале предложения, мы стараемся найти тексты, где наш запрос является ремой.

Также при использовании фактора «пар слов» (в документе ищутся пары, состоящие из слов из запроса) имеет смысл искать «все слова со всеми», поскольку в запросах слова необязательно зависят друг от друга линейно: к примеру, в запросе *блюда из молодой свинины* будет правильным выделить такие словосочетания: *блюда из свинины*, *молодая свинина* – как видим, слова не связаны друг с другом только линейно, т.к. слово *молодая* относится исключительно к свинине, а со словами *блюда из* сочетается опять же слово *свинина*, а не *молодая* (т.е. *блюда из свинины*, а не *блюда из молодой*). Подобный подход был применен в [1].

Параметр учета близости слов к началу предложения является частью нашего поискового алгоритма и дает определенный прирост качества – соответственно, мы уже сейчас успешно используем лингвистические данные при поиске документов.

Таким образом, языковой анализ запросов и текстов можно использовать для совершенствования поискового алгоритма не только тематических, научных, но и больших поисковых систем в Интернете, однако лингвистический анализ следует не буквально включать в поисковый алгоритм, а использовать его для выявления новых закономерностей, которые можно было бы описать статистически.

Мы планируем в будущем уделять внимание синтаксическому и дискурсивному анализу запросов и текстов в коллекции, но, безусловно, важной задачей останется сохранить промышленные характеристики поискового алгоритма.

V. Заключение

Эксперименты в рамках семинара РОМИП показали, что применение кворума и оптимизация параметров новым методом обеспечивают хорошее качество поисковой выдачи. Высокий уровень оценок по всем коллекциям говорит о стабильности нашего поискового алгоритма.

Литература

[1]. Агеев М.С., Добров Б.В., Красильников П.В., Лукашевич Н.В., Павлов А.М., Сидоров А.В., Штернов С.В. УИС РОССИЯ в РОМИП 2007: поиск и классификация. // Труды РОМИП 2007-2008. Санкт-Петербург: НУ ЦСИ, 2008

[2]. Зорин В.М, Копченова Н.В 1993. Некоторые методы решения оптимизационных задач. М:Издательство МЭИ.

[3]. Сафронов А.В. HeadHunter на РОМИП-2008. // Труды РОМИП 2007-2008. Санкт-Петербург: НУ ЦСИ, 2008

[4]. Сегалович И., Маслов, М. Яндекс на РОМИП-2004. Некоторые аспекты полнотекстового поиска и ранжирования в Яндексе. // РОМИП'2004: Тр. конф. - М., 2004

[5]. Татевосян, С., Брызгалова, Н. KM.RU на РОМИП-2008. Оптимизация параметров поискового алгоритма. // Труды РОМИП 2007-2008. Санкт-Петербург: НУ ЦСИ, 2008

KM.RU at RIRES-2009

S. Tatevosyan, N. Bryzgalova

The present article is devoted to participation of KM.RU Search at RIRES-2009. We give a brief description of our modified information retrieval algorithm and talk about new optimization methods. The paper also reports on experimental results obtained at RIRES-2009.